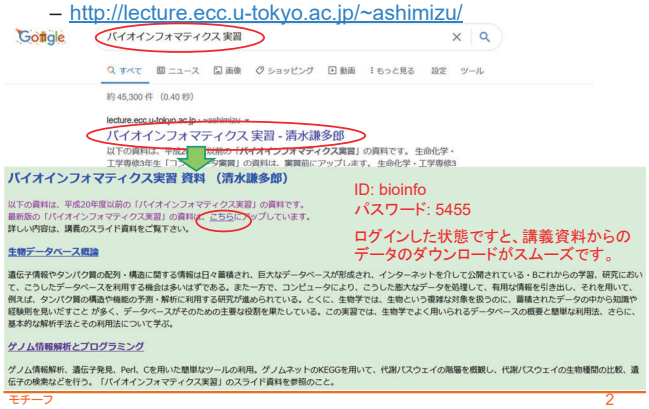
アミノ酸配列からタンパク質の機能を探る ファミリー、ドメイン、モチーフ

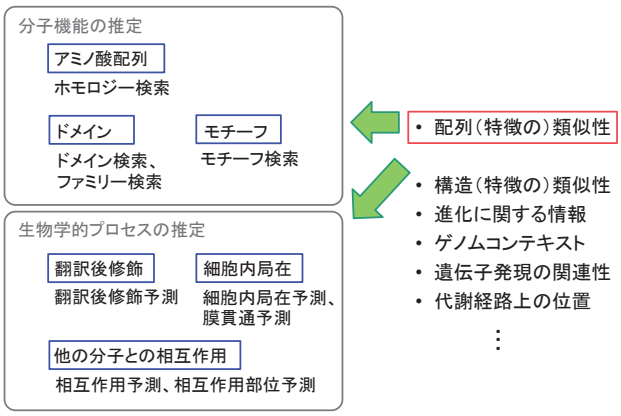
清水謙多郎
shimizu@bi.a.u-tokyo.ac.jp

実習の資料について

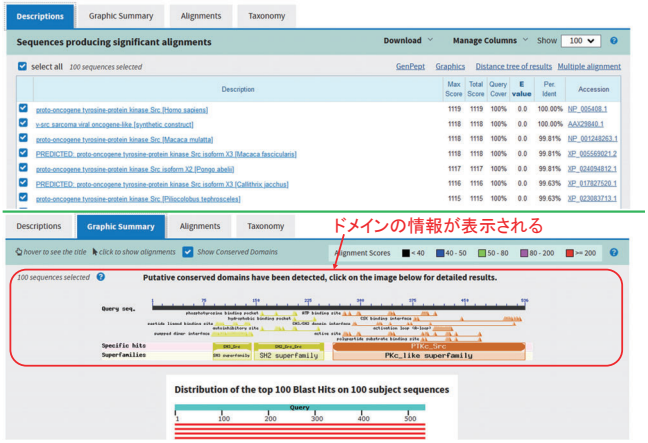
- 「バイオインフォマティクス 実習」で検索
  - http://lecture.ecc.u-tokyo.ac.jp/~ashimizu/



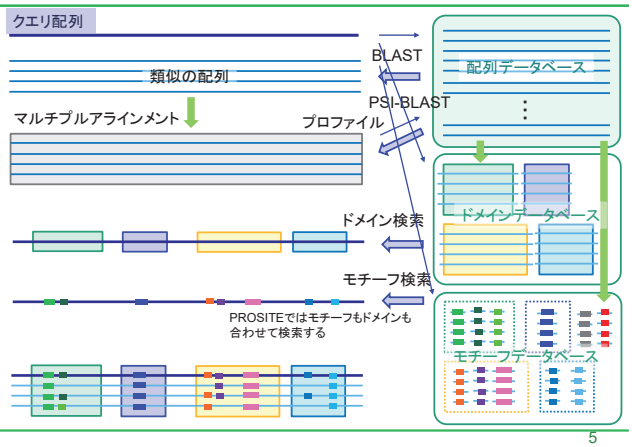
タンパク質の機能予測



BLAST検索の結果



タンパク質の配列からの機能予測

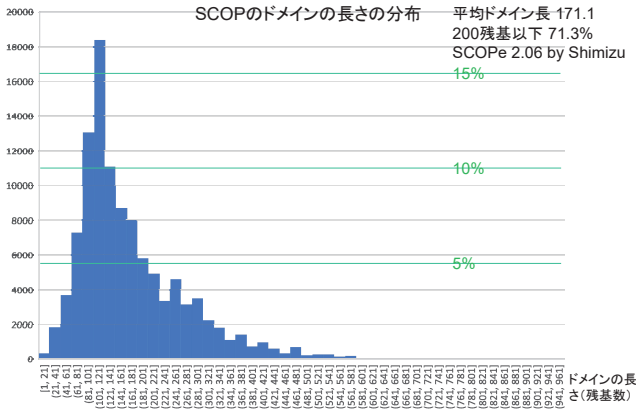


タンパク質のドメイン

- ドメイン（domain）：タンパク質構造を構成する構造的あるいは機能的にまとまった単位
  - タンパク質は1つまたは複数のドメインから構成される
  - 多くは200残基程度からそれ以下
    - G. A. Petsko, D. Ringe, Protein Structure and Function, New Science Press, 2004.
- 各ドメインには、共通した配列・構造特徴および機能をもつものがあり、それらは「同じドメイン」とみなされる
- 「同じドメイン」が複数のタンパク質に現れることが多い
- 複数のドメインから構成されるマルチドメインタンパク質では、各ドメインに対応する機能を合わせもつことが多い
  → ドメインの構成によって、タンパク質の機能が決まる

参考

タンパク質のドメイン

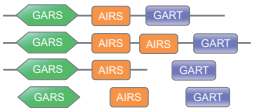


参考

タンパク質のドメイン

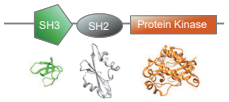
- 真核生物のタンパク質の65%、原核生物のタンパク質の40%がマルチドメイン構成である
  - Ekman D, et al. "Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions". Journal of Molecular Biology. **348**: 231-243 (2005). doi:10.1016/j.jmb.2005.02.007.
- ドメインは、それ自身、独立してフォールドできるような構造で、真核生物のマルチドメインタンパク質の1つのドメインが、原核生物で単独のタンパク質で存在する例が見られる
  - Davidson JN, Chen KC, Jamison RS, Musmanno LA, Kern CB (March 1993). "The evolutionary history of the first three enzymes in pyrimidine biosynthesis". BioEssays. **15**: 157–164. doi:10.1002/bies.950150303.

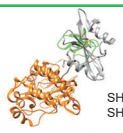
三機能プリン生合成タンパク質アデノシン-3
GAR合成酵素（GARS）、AIR合成酵素（AIRS）、GARホルミル基転移酵素（GART）



ドメインの例

プロテインチロシンキナーゼ Src

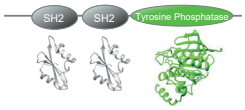


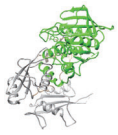


UniProtKB P12931
PDB 2h8h

SH2: Src homology 2 domain
SH3: Src homology 3 domain

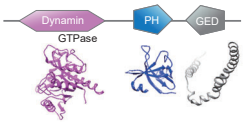
プロテインチロシンホスファターゼ

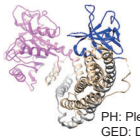




UniProtKB Q06124
PDB 2shp

ダイナミン



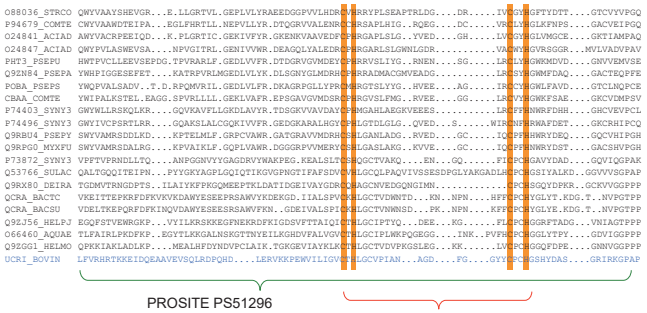


UniProtKB Q9UQ16
PDB 5a3f

PH: Pleckstrin homology domain
GED: Dynamin GTPase effector domain

Rieske型鉄硫黄クラスター結合部位

Rieske型鉄硫黄クラスター結合部位の配列アラインメント



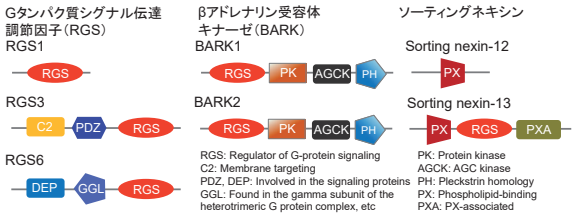
PROSITE PS51296
Rieske [2Fe-2S] 鉄硫黄ドメイン
のプロファイル

C-x-H-x(15-17)-C-x-x-H
鉄結合部位のモチーフ

ドメイン全体にわたる配列特徴　モチーフの局所的な配列特徴

ファミリーとドメイン

- タンパク質ファミリー
  - 進化的類縁関係をもつタンパク質のグループ
  - 共通の機能をもつタンパク質のグループ
- 「ドメインの構成によって、タンパク質の機能が決まる」
  - ファミリーはドメインの構成によって決まる
  - 同じドメインが異なるファミリーに現れることがある

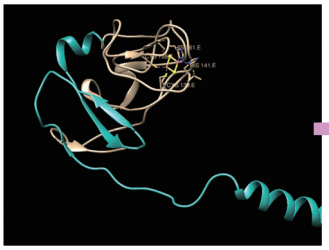


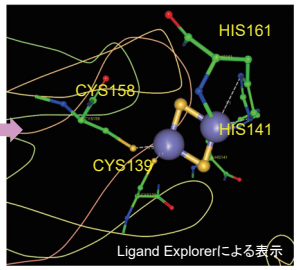
RGS: Regulator of G-protein signaling
C2: Membrane targeting
PDZ, DEP: Involved in the signaling proteins
GGL: Found in the gamma subunit of the heterotrimeric G protein complex, etc

PK: Protein kinase
AGCK: AGC kinase
PH: Pleckstrin homology
PX: Phospholipid-binding
PXA: PX-associated

鉄硫黄クラスター周辺の構造

Cytochrome bc1（UniProtKB: UCRI_BOVIN, PDB: 1L0L Eチェイン）
鉄硫黄クラスター(2Fe-2S)周辺の構造





結合部位（結合サイト）

ファミリーとドメイン

- タンパク質の機能を探るのに、構成するドメインを同定することは重要
  - ドメインとその機能、タンパク質が構成するドメインをデータベースに登録
  - 配列の類似性を手がかりとして、ドメインを同定し、機能を推定する
    - ドメインを同定することは、タンパク質のファミリーを同定することにつながる

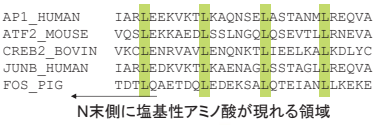
ドメイン、ファミリーのデータベースと高感度の検索

Leucine Zipper

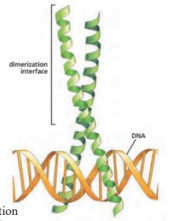
DNAに結合する転写因子に見られる、αヘリックスの2量体

L-x(6)-L-x(6)-L-x(6)-L

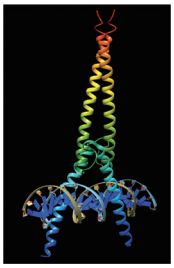
PROSITE PS00029



N末側に塩基性アミノ酸が現れる領域



Molecular Biology of the Cell, 6 edition



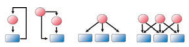
OmoMYC bound to double-stranded DNA
PDBID: 5l50

モチーフ

- 複数の塩基配列やアミノ酸配列に共通に見られる、保存された(短い)配列パターン
  - 特定の機能、性質に関係する
    - 塩基配列 → 転写因子結合部位、TATAボックス、スプライシング部位(GT-AGモチーフ)など
- アミノ酸配列のモチーフ
  - 酵素の活性部位、他の分子との相互作用部位、翻訳後修飾、ドメインを特徴づける配列パターンなどに関係
  - (配列全体の類似度が低くても)特定の機能、性質に関係する部分は強く保存されている傾向にある

- タンパク質の立体構造 → 構造モチーフ

- ネットワーク → ネットワークモチーフ

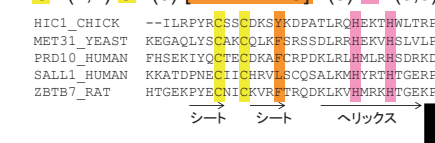


Zinc Finger

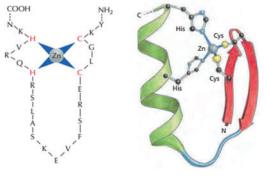
DNAに結合する転写因子に見られる

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

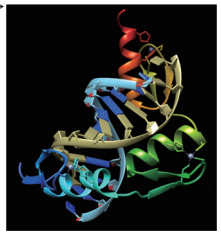
PROSITE パターン PS00028
プロファイル PS50157



シート　シート　ヘリックス



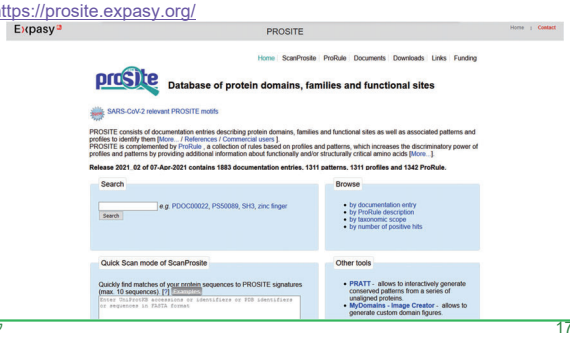
Molecular Biology of the Cell, 6 edition



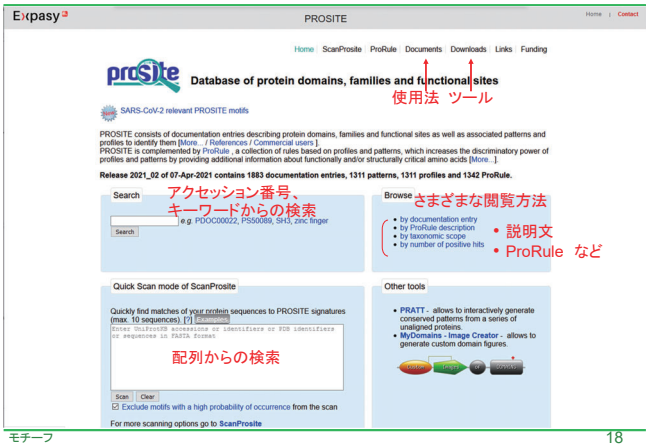
Zif268-DNA complex, PDBID:1zaa

PROSITEデータベース

- タンパク質のファミリー、ドメイン、機能部位のデータベース
- それらを特徴づける配列パターン（モチーフ）を登録し、その検索機能をもつ
- ファミリーに特徴的な機能やドメインの構造を記述したドキュメンテーションエントリーと、モチーフの表現を登録したエントリーをもつ
- EBIのExPaSy Proteomics Serverのデータベースの一つ
- https://prosite.expasy.org/

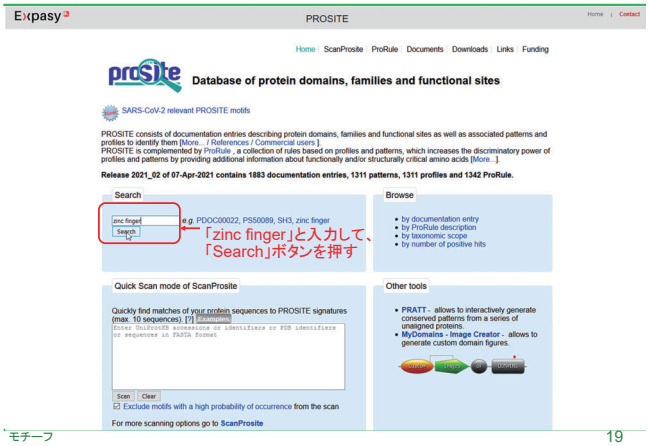


PROSITEデータベースの利用



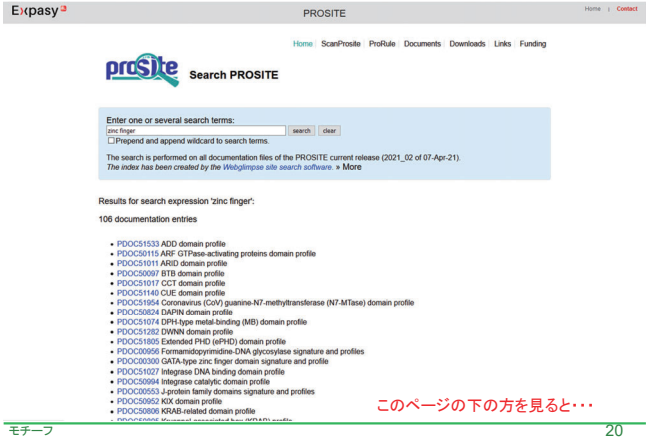
参考

PROSITEの検索（1）



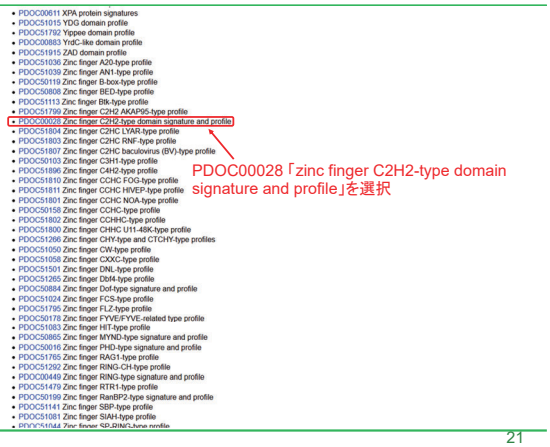
参考

PROSITEの検索（2）



参考

PROSITEの検索（3）



参考

PROSITEの検索（4）

参考

PROSITEの検索（5）

参考

PROSITEの検索（6）

参考

PROSITEの検索（7）

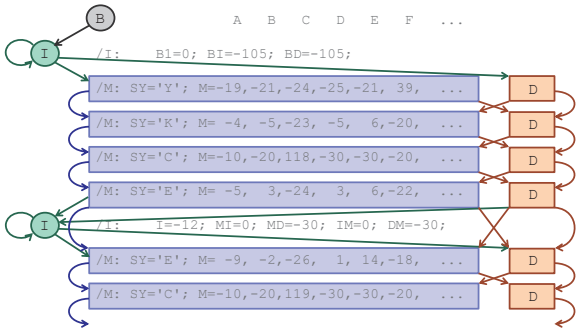
プロファイルの実際

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTUWYZ'; LENGTH=28;
/DISJOINT: DEFINITION=PROTECT; N1=3; N2=26;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=-0.6689000; R2=0.0207831; TEXT='NScore';
/NORMALIZATION: MODE=-1; FUNCTION=LINEAR; R1=1259.5933838; R2=1.0076727; TEXT='Heuristic 5.0%';
/CUT_OFF: LEVEL=0; SCORE=442; H_SCORE=1705; N_SCORE=8.5; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=345; H_SCORE=1607; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: D=-20; I=-20; B1=-50; E1=-50; MI=-105; MD=-105; IM=-105; DM=-105;

            A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
/I:         B1=0; BI=-105; BD=-105;
/M: SY='Y'; M=-19,-21,-24,-25,-21, 39,-28, 10, -2,-17,  2,  0,-17,-28,-19,-13,-18,-10, -7, 12, 48,-21;
/M: SY='K'; M= -4, -5,-23, -5,  6,-20,-18, -9,-14,  7,-15, -7, -5, -9,  3,  2, -2, -2, -9,-25,-12,  4;
/M: SY='C'; M=-10,-20,118,-30,-30,-30,-30,-30,-20,-20,-20,-40,-30,-30,-30,-10,-10,-10,-10,-50,-28,-30;
/M: SY='E'; M= -5,  3,-24,  3,  6,-22,-11, -6,-20,  1,-21,-14,  4, -1,  1, -3,  5,  2,-18,-29,-15,  3;
/I:         I=-12; MI=0; MD=-30; IM=0; DM=-30;
/M: SY='E'; M= -9, -2,-26,  1, 14,-18,-17, -4,-13, -1,-11, -8, -5,-12,  4, -5, -5, -8,-12,-24, -9,  8;
/M: SY='C'; M=-10,-20,118,-30,-30,-30,-30,-30,-20,-20,-20,-40,-30,-30,-30,-10,-10,-10,-10,-50,-28,-30;
/M: SY='G'; M= -3, -1,-28, -1, -7,-28, 36,-11,-33,-11,-27,-18,  4,-15,-10,-12,  1,-13,-27,-24,-23, -9;
/M: SY='K'; M=-10, -2,-28, -3,  8,-25,-19, -7,-26, 36,-24, -8, -1,-12, 10, 27, -9, -9,-18,-19, -8,  8;
/M: SY='L'; M=  8, -7, -9,-11, -7,-17, -7,-14,-16, -8,-16, 11, -4,-15, -6, -5,  8,  4, -7,-27,-15, -7;
/M: SY='F'; M=-19,-29,-19,-37,-28,117,-29,-17,  0,-28,  0,-28,-30,-36,-19,-19, -9, -1,  1,117,-28;
/M: SY='S'; M=  0, -5,-17, -9, -6,-16,-11,-10,-14, -3,-16,-10,  0,-12, -4,  0,  8,  7, -8,-27,-12, -6;
/M: SY='R'; M=-10, -3,-20, -4,  0,-18,-17,  2,-19,  3,-16, -8,  0,-17,  8,  9, -1, -3,-17,-19, -5,  3;
/M: SY='R'; M= -4, -4,-22, -6,  0,-19,-13, -5,-18,  7,-18, -9,  1,-10,  2,  8,  2, -2,-14,-25,-11,  0;
/M: SY='S'; M=  2, -1,-16, -1, -1,-18, -4, -6,-19, -7,-22,-14,  4,-12, -2, -7, 16,  7,-13,-29,-12, -2;
/M: SY='N'; M= -5,  5,-20,  1,  0,-18,-10,  8,-18, -5,-18,-11,  9,-16,  1, -4,  4, -1,-17,-27, -7, -1;
/M: SY='L'; M=-11,-29,-20,-30,-20, 12,-29,-19, 17,-27, 43, 18,-28,-29,-20,-18,-28,-10,  9,-18,  2,-20;
/M: SY='R'; M= -6, -6,-22,-10, -4,-15,-20, -8, -7,  2, -9, -3, -2,-16,  0,  3, -3,  0, -6,-24, -8, -3;
/M: SY='R'; M= -6, -6,-23, -7,  0,-19,-17, -7,-14,  7,-13, -6, -2,-16,  4, 12, -3, -3,-10,-24,-10,  0;
/M: SY='R'; M=-20, 10,-30,  0,  0,-26,-20,-30,-10,-20,  0, 10,-20, 10,  0,-10,-20,-30,-30, 20,  0;
/M: SY='Q'; M=-10,-10,-25,-12,  1,-16,-22, -2, -6,  1, -3,  6, -9,-17, 13,  3, -9, -8, -9,-19, -4,  6;
/M: SY='R'; M=-13, -8,-26, -9,  0,-19,-19, -4,-21, 20,-16, -6, -2,-17,  6, 35, -8, -7,-14,-21, -9,  0;
/I:         I=-12; MI=0; MD=-29; IM=0; DM=-29;
/M: SY='V'; M= -3,-16,-17,-21,-17, -6,-25,-20, 11,-15,  2,  3,-12,-18,-14, -2,  9, 13,-25, -7,-17;
/M: SY='R'; M=-20,  0,-30,  0,  0,-26,-20,-30,-10,-20,  0, 10,-20, 10,  0,-10,-20,-30,-30, 19,  0;
/M: SY='T'; M=  1, -2,-13, -8, -6,-14,-15,-15,-12, -7,-13, -9,  0,-12, -6, -8, 14, 25, -5,-29,-12, -6;
/M: SY='G'; M= -3, -4,-27, -4, -8,-27, 34,-14,-32, -9,-25,-16,  2,-15,-10, -9,  0,-13,-24,-23,-23,-10;
/M: SY='E'; M= -9,  6,-27, 12, 33,-26,-17, -3,-24,  7,-18,-15, -1, -6, 12,  0, -2, -8,-22,-28,-16, 22;
/M: SY='K'; M=-11, -1,-28, -2,  6,-25,-17, -6,-27, 32,-25,-10,  1,-11,  8, 28, -7, -8,-19,-22,-11,  6;
/M: SY='P'; M= -7,-14,-32, -9, -1,-24,-17,-15,-18, -7,-23,-15,-13, 51, -7,-13, -6, -6,-23,-28,-22, -7;
/I:         E1=0;
```


参考

PROSITEのGeneralized Profile

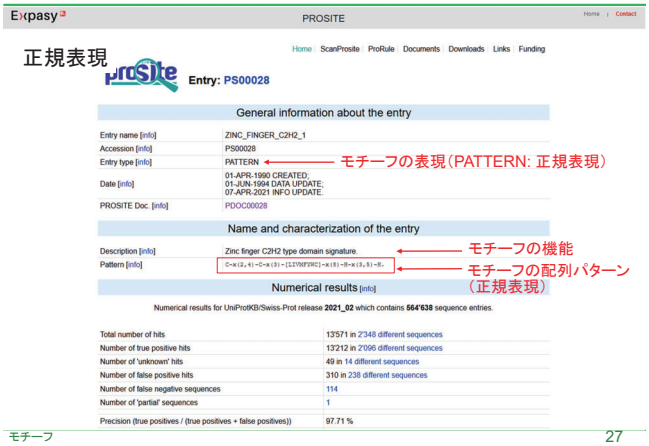
状態遷移に基づく記法
挿入、欠失に対応し、隠れマルコフモデルと同等の機能をもつ



参考

PROSITEの検索（8）

正規表現



参考

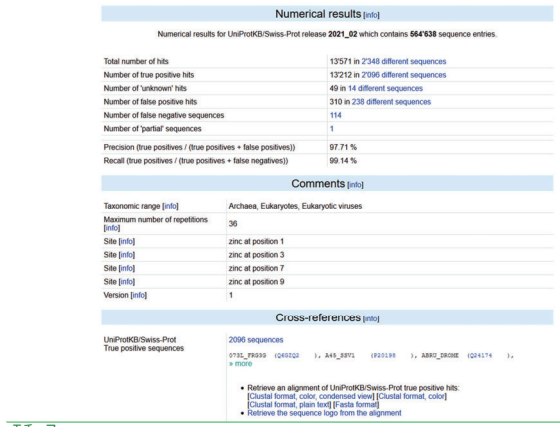
PROSITEの正規表現

- The standard IUPAC one-letter codes for the amino acids are used.
- The symbol 'x' is used for a position where any amino acid is accepted.
- Ambiguities are indicated by listing the acceptable amino acids for a given position, between square parentheses '[ ]'.
  → [ALT] stands for Ala or Leu or Thr.
- Ambiguities are also indicated by listing between a pair of curly brackets '{ }' the amino acids that are not accepted at a given position.
  → {AM} stands for any amino acid except Ala and Met.
- Each element in a pattern is separated from its neighbor by a '-'.
- Repetition of an element of the pattern can be indicated by following that element with a numerical value or a numerical range between parenthesis.
  → x(3) corresponds to x-x-x, x(2,4) corresponds to x-x or x-x-x or x-x-x-x.
- When a pattern is restricted to either the N- or C-terminal of a sequence, that pattern either starts with a '<' symbol or respectively ends with a '>' symbol. In some rare cases (e.g. PS00267 or PS00539), '>' can also occur inside square brackets for the C-terminal element. 'F-[GSTV]-P-R-L-[G>]' means that either 'F-[GSTV]-P-R-L-G' or 'F-[GSTV]-P-R-L>' are considered.
- A period ends the pattern.

DocumentsのDocument describing the syntax of profiles in PROSITE（A generalised profile syntax for protein and nucleic acid sequence motifs, profile.txt）に記載

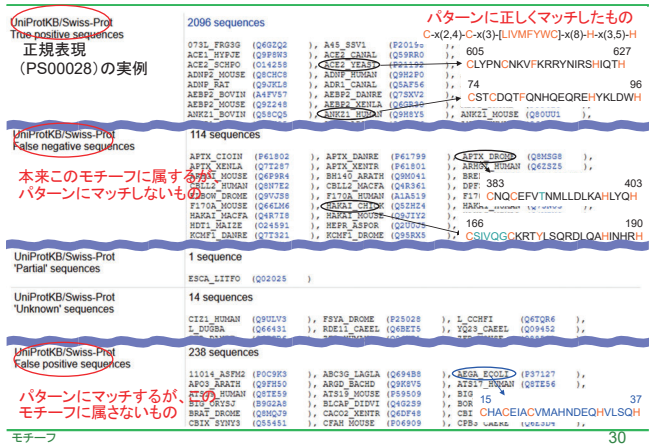
参考

PROSITEの検索（9）



参考

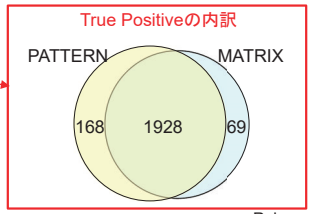
PROSITEの検索（10）



参考

Zinc Finger C2H2-typeの事例

	PATTERN（PS00028）	MATRIX（PS50157）
True Positive	2096	1997
False Negative（検討中のもの）	114	210
False Positive（検討中のもの）	238	9

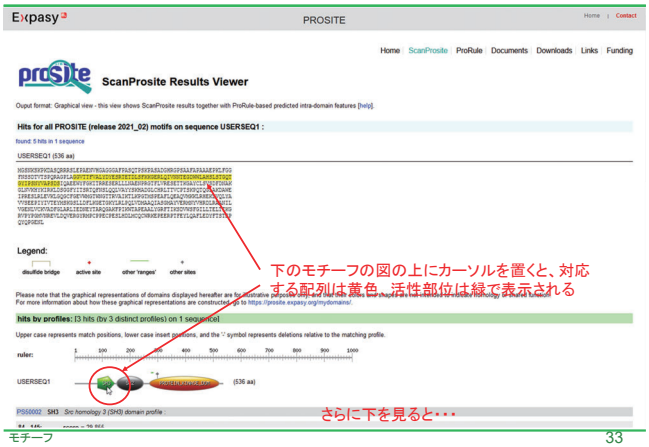


Release 2021_02 of 07-Apr-2021

参考

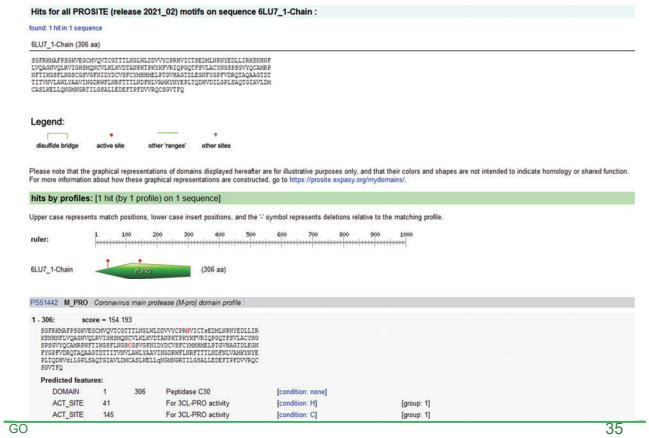
PROSITEの検索（11）

PROSITEの検索（12）

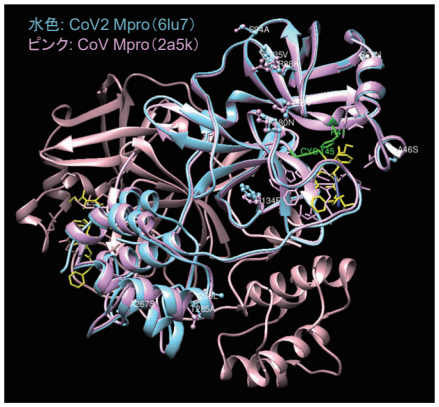


PROSITEの検索（13）

PROSITEの検索（14）



SARS CoV2とSARS CoVの比較（1）



SARS CoV2とSARS CoVの比較（2）

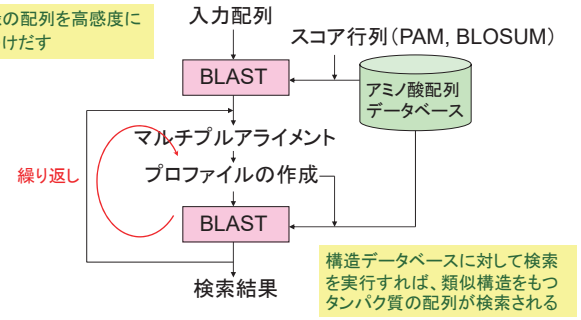
```
CoV2_Mpro SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGH
2a5k.1.A  SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDTVYCPRHVICTAEDMLNPNYEDLLIRKSNHSFLVQAGNVQLRVIGH
                                          35   41   46                65
CoV2_Mpro SMQNCVLKLKVDTANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFC
2a5k.1.A  SMQNCLLRLKVDTSNPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNHTIKGSFLNGSCGSVGFNIDYDCVSFC
           86 88  94                                      134      145
CoV2_Mpro YMHHMELPTGVHAGTDLEGNFYGPFVDRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYE
2a5k.1.A  YMHHMELPTGVHAGTDLEGKFYGPFVDRQTAQAAGTDTTITLNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYE
                                          202
CoV2_Mpro PLTQDHVDILGPLSAQTGIAVLDMCASLKELLQNGMNGRTILGSALLEDEFTPFDVVRQCSGVTFQ
2a5k.1.A  PLTQDHVDILGPLSAQTGIAVLDMCAALKELLQNGMNGRTILGSTILEDEFTPFDVVRQCSGVTFQ
                                  267      285 286
```


PSI-BLAST

Position-Specific Iterated BLAST

位置依存スコア　繰り返し

遠縁の配列を高感度に見つけだす



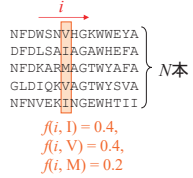
構造データベースに対して検索を実行すれば、類似構造をもつタンパク質の配列が検索される

参考

PSI-BLASTのプロファイル

- 配列のマルチプルアラインメントの各位置（カラム）i において、
  - $n(i,j)$: その位置のアミノ酸 j の出現度数
  - その位置のアミノ酸 j の出現頻度 $f(i,j)$ は、

$$f(i,j)=\frac{n(i,j)}{\sum_j n(i,j)}$$



$f(i, \mathrm{I}) = 0.4,$
$f(i, \mathrm{V}) = 0.4,$
$f(i, \mathrm{M}) = 0.2$

- PSI-BLASTのプロファイル$S(i,j)$の計算

$$S(i,j)=\log\frac{Q(i,j)}{P(j)}$$

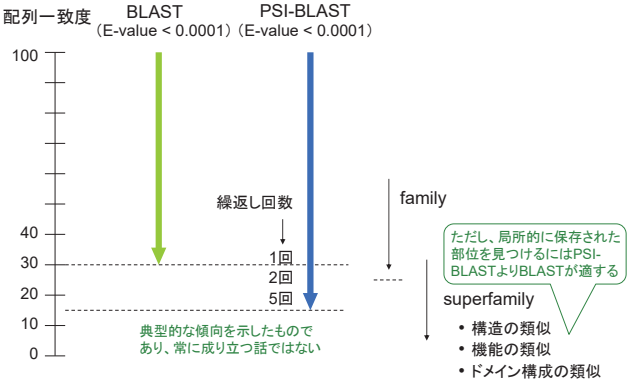
  - $P(j)$はバックグラウンドでのアミノ酸jの出現頻度
  - $Q(i,j)$は、その位置に現れると見積もられるアミノ酸 j の出現頻度

$$Q(i,j)=\frac{\alpha f(i,j)+\beta g(i,j)}{\alpha+\beta}$$

  - αは、そのカラムに出現するギャップも含めたアミノ酸の種類の数−1、β = 10
  - $q(i,j)$をアミノ酸の置換スコアとすると、

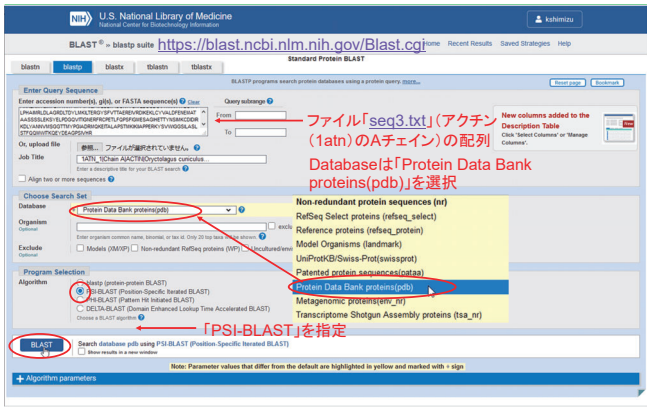
$$g(i,j)=\sum_k \frac{f(i,k)}{P(k)}q(k,j)$$

PSI-BLASTによる検索の基準



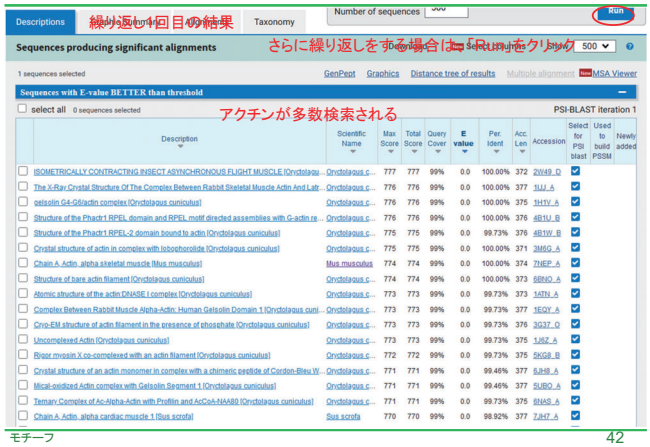
参考

PSI-BLASTの利用（1）



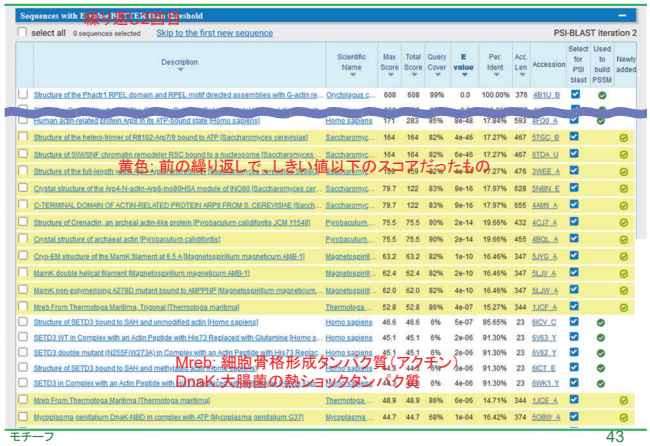
参考

PSI-BLASTの利用（2）



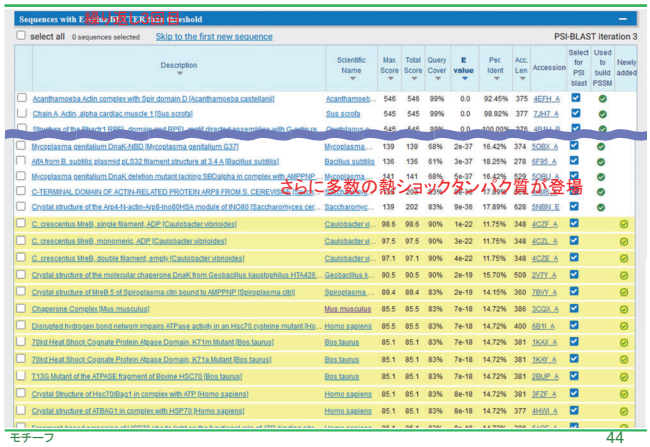
参考

PSI-BLASTの利用（3）



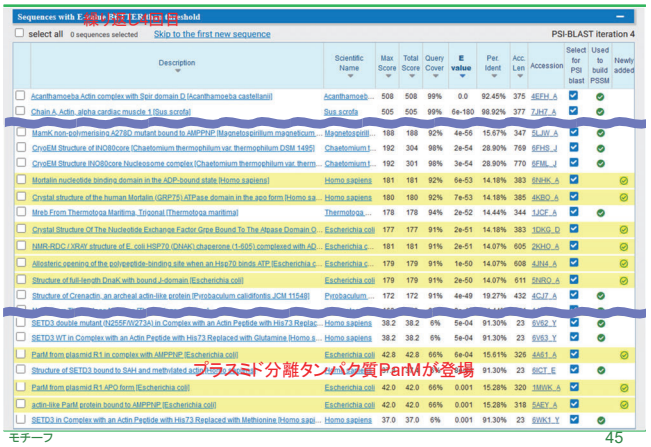
参考

PSI-BLASTの利用（4）



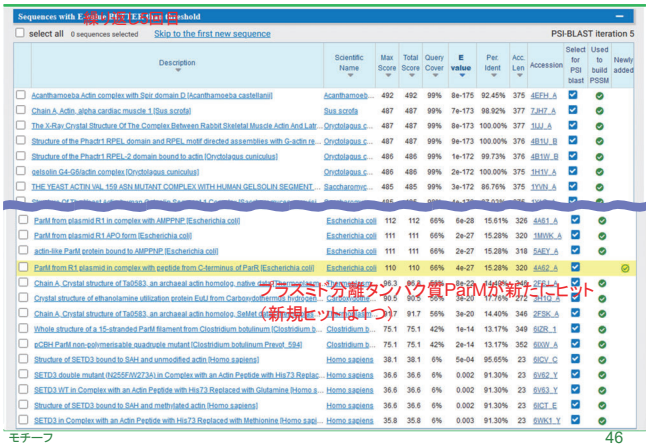
参考

PSI-BLASTの利用（5）



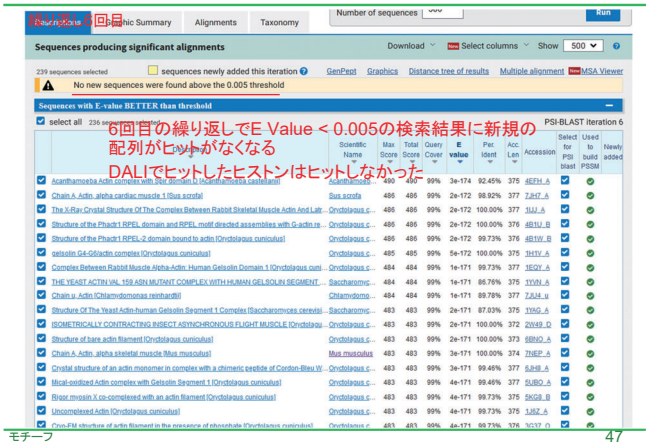
参考

PSI-BLASTの利用（6）



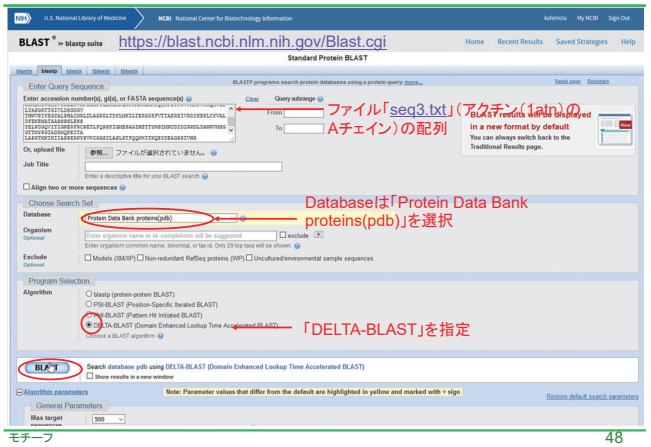
参考

PSI-BLASTの利用（7）



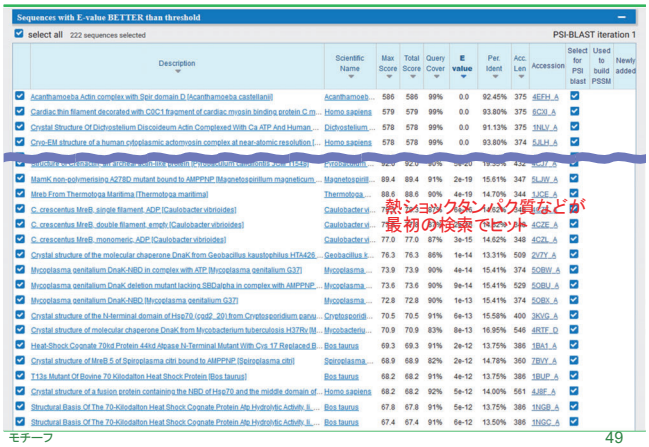
参考

DELTA-BLASTの利用（1）



参考

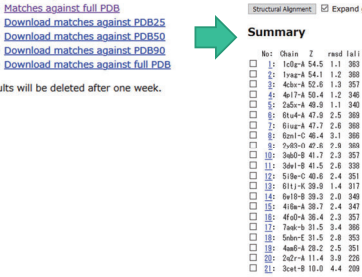
DELTA-BLASTの利用（2）



熱ショックタンパク質などが最初の検索でヒット

参考

DALIの検索結果の例（2）



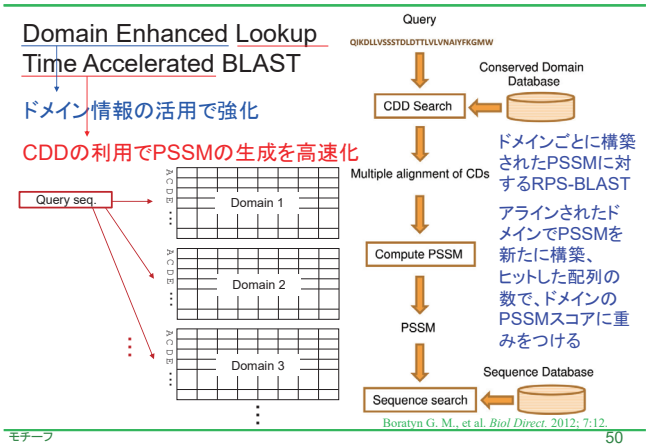
参考

DELTA-BLAST

Domain Enhanced Lookup Time Accelerated BLAST

ドメイン情報の活用で強化

CDDの利用でPSSMの生成を高速化



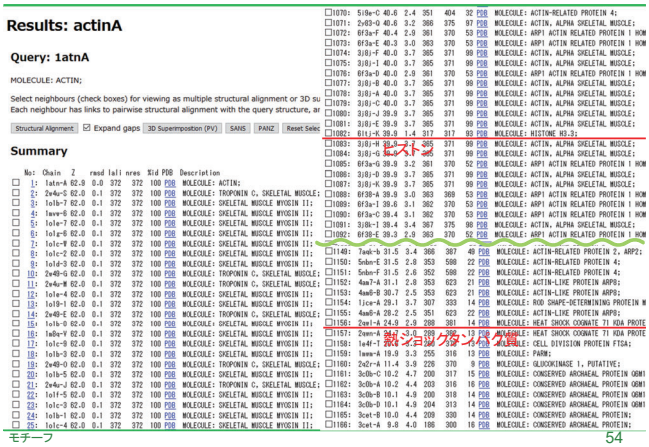
ドメインごとに構築されたPSSMに対するRPS-BLAST

アラインされたドメインでPSSMを新たに構築、ヒットした配列の数で、ドメインのPSSMスコアに重みをつける

Boratyn G. M., et al. *Biol Direct.* 2012; 7:12.

参考

DALIの検索結果の例（3）



参考

DALI

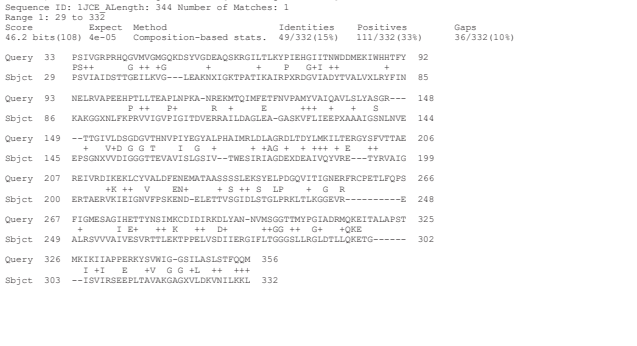
- DALI: タンパク質の構造比較の手法およびデータベース
  - http://ekhidna2.biocenter.helsinki.fi/dali/

- PDB search
  - 類似構造をPDB全体に対して検索するサービス
- All against allの検索
  - 全PDB構造の間で比較した結果を格納したデータベース
- Pairwise
  - ペアワイズで構造を比較

参考

PSI-BLASTの結果の例

1jce Aチェイン（Mreb From Thermotoga Maritima）
PSI-BLASTの2回目の繰り返しで登場



参考

DALIの検索結果の例（1）

構造比較のみで、似た構造を検索

PDB ID：アクチンlatnのチェインAを入力（seq3.txt）

ジョブ名

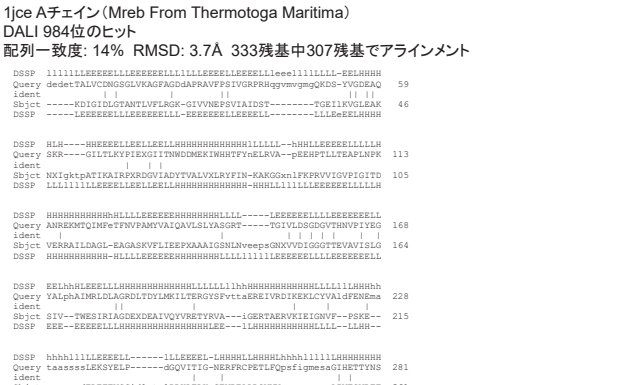
メールアドレスを入力

「submit」をクリック

参考

DALIの結果の例（1）

1jce Aチェイン（Mreb From Thermotoga Maritima）
DALI 984位のヒット
配列一致度: 14%　RMSD: 3.7Å　333残基中307残基でアラインメント



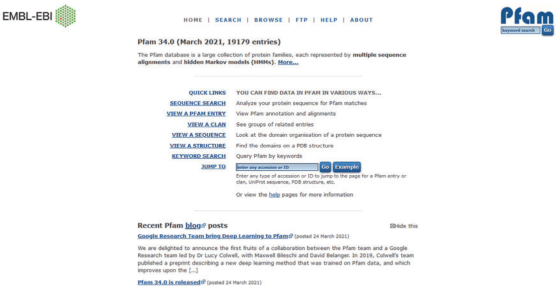
参考

DALIの結果の例（2）

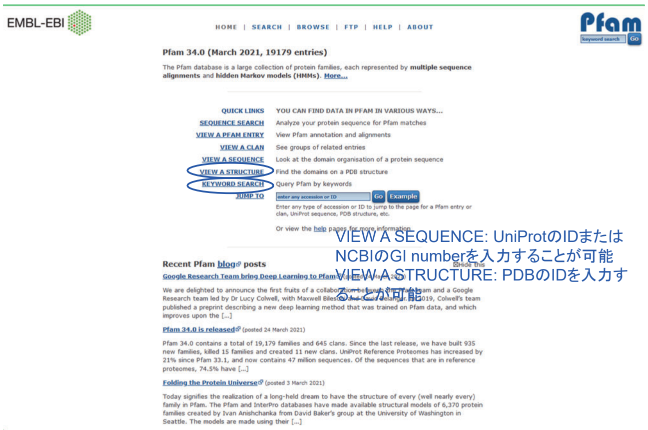
```
DSSP  HLLLLLLLHHHHH-LLEEEELHHHHLLLHHHHHHHHHHlllllllLLEELLLLHHHHHH
Query IMKCDIDIRKDLY-ANNVMSGGTTMYPGIADRMQKEITalapstmkIKIIAPPERKYSVW  340
ident   |      |       ||    |    |||        |  |   |    |
Sbjct LEKTPPELVSDIIeRGIFLTGGGSLLRGLDTLLQKETG--------ISVIRSEEPLTAVA  313
DSSP  HHLLLHHHHHHHHhHLEEEELHHHLLLLHHHHHHHHHL--------LLEEELLLLLHHH

DSSP  HHHHHHHHLLL-LHHHLEEHhhhhhllhhhhl
Query IGGSILASLST-FQQMWITKqeydeagpsivhr  372
ident  |
Sbjct KGAGXVLDKVNiLKKLQGAG-------------  333
DSSP  HHHHHHHHLHHhHLLLLEEL-------------
```


Pfam

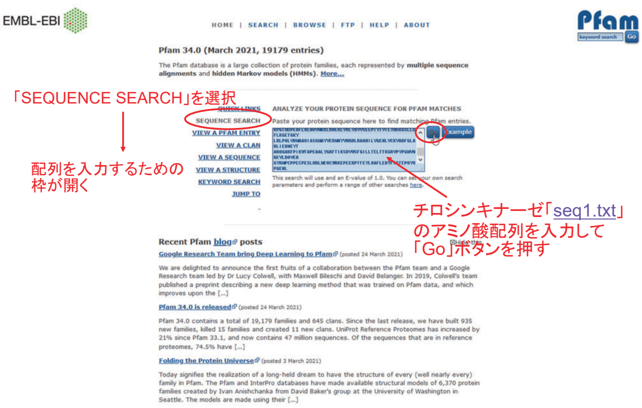
- Protein families database of alignments and HMMs
- タンパク質のドメインとファミリー、それらを特徴づける配列パターンを表す隠れマルコフモデル（HMM）を登録
- ドメインの立体構造も表示（そのドメインをもつタンパク質の構造）
- 人手によるアノテーション
- https://pfam.xfam.org/



Pfamの検索（1）

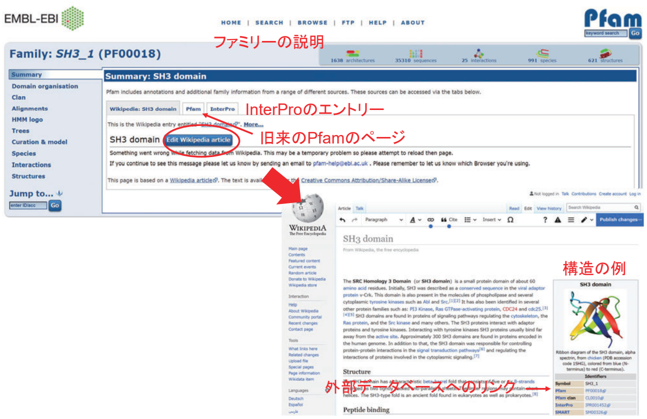


Pfamの検索（2）

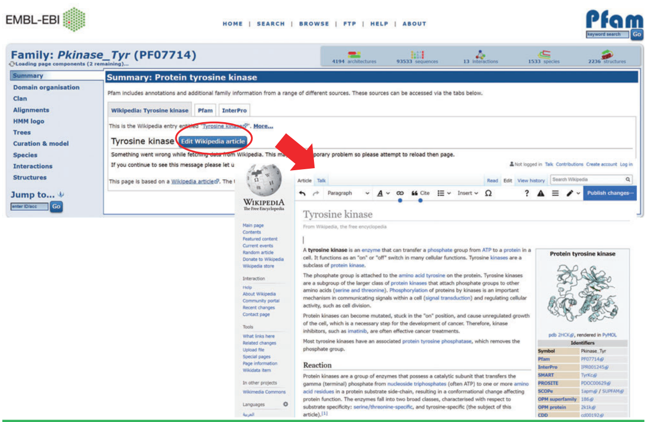


Pfamの検索（3）

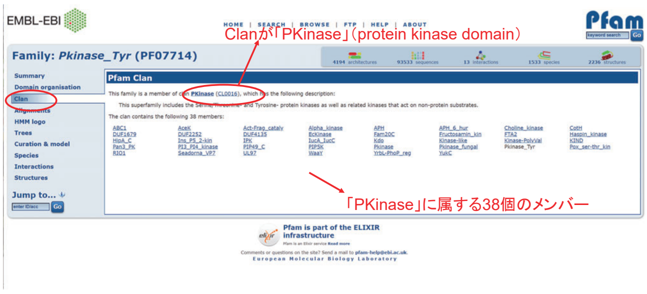
Pfamの検索（4）



Pfamの検索（5）

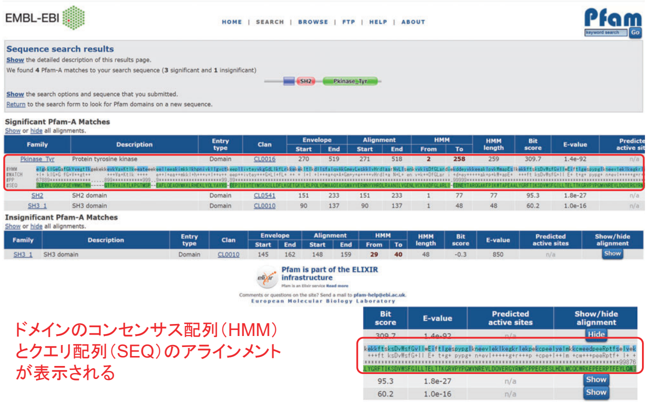


Pfamの検索（6）



Pfamの検索（7）

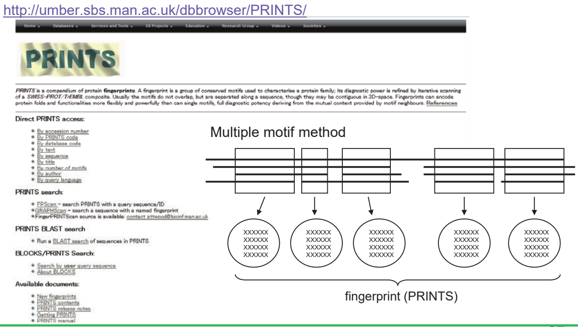
Pfamの検索（8）



参考

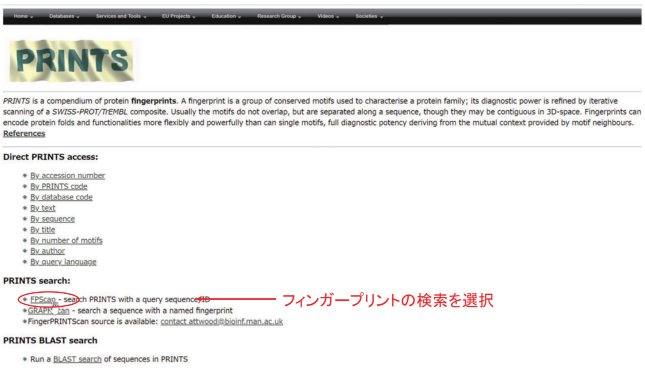
PRINTS

- マルチプルアラインメントから最も保存された領域（モチーフ）を抽出
- これらのモチーフをもとに、SWISS-PROTおよびTrEMBLの配列に対し探索をかけ、これを収束するまで繰り返して得られたもの
- モチーフの抽出、アノテーションは人手による
- モチーフ: BLOCKSと同じ、ギャップなしで保存された領域（抽出の手法は異なる）
- 1つのタンパク質ファミリーを複数のモチーフ（フィンガープリント）で記載
- http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/



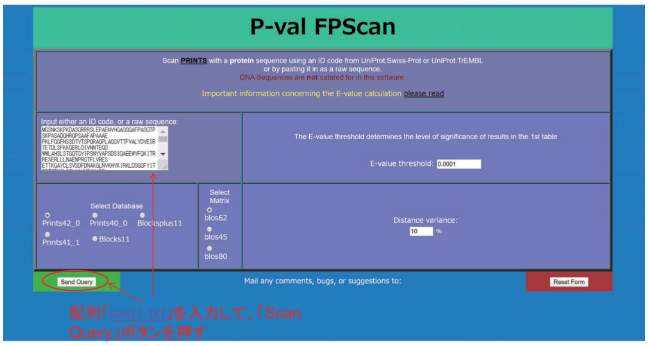
参考

PRINTSの検索（1）



参考

PRINTSの検索（2）



参考

PRINTSの検索（3）

PRINTS42_0 and matrix blos62

Scan of sequence: USER_SEQUENCE

Highest scoring fingerprints for your query			
Fingerprint	E-value	GRAPHScan	Motif3D
TYRKINASE (relations)	7.881181e-39	Graphic	
SH2DOMAIN (relations)	1.155504e-24	Graphic	
SH3DOMAIN (relations)	4.669176e-14	Graphic	

for further information choose any of the following options

- Simple - Top Ten
- Detailed - Top Ten (detailed by motif)

Back to top

Ten top scoring fingerprints for your query

Ancestry	Fingerprint	No. of Motifs	SumId	AveId	PfScore	Pvalue	Evalue	GRAPHScan	
TYRKINASE	TYRKINASE	5 of 5	3.1e+02	61	3118	8.4e-45	7.9e-39	IIIII	Graphic
SH2DOMAIN	SH2DOMAIN	5 of 5	289.25	57.85	1859	8.8e-31	1.2e-24	IIIII	Graphic
SH3DOMAIN	SH3DOMAIN	4 of 4	178.81	44.70	1203	2.8e-20	4.7e-14	IIII	Graphic
SPECTRNALPHA	SPECTRNALPHA	2 of 3	68.40	34.20	641	1.4e-07	0.33	iI.	Graphic
GLHYDRLASE15	GLHYDRLASE15	2 of 7	59.45	29.73	399	0.00041	1e+03	I.....i	Graphic
ADENOVSFIBRE	ADENOVSFIBRE	2 of 7	68.16	34.08	306	0.0005	1.6e+03	..iI...	Graphic
7TM-->GPCRCLAN-->GPCRRHODOPSN-->OREXINR	OREXINR	2 of 10	60.71	30.36	430	0.0012	2.8e+03	......i..I	Graphic
MEROZOITESA	MEROZOITESA	2 of 15	49.39	24.70	399	0.0039	6.8e+03	i.........i....	Graphic

参考

PRINTSの検索（4）

Ten top scoring fingerprints for your query. Detailed by motif

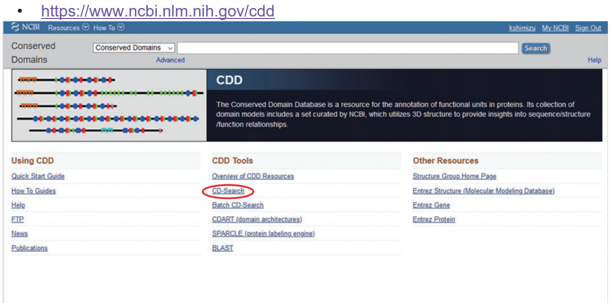
FingerPrint Name	Motif Number	IdScore	PfScore	Pval	Sequence	Length	low	Pos	high
TYRKINASE	1 of 5	58.00	405	4.58e-08	TEYMSKGSLLDFLK	14	0	341	0
	2 of 5	62.00	628	2.92e-10	YVERMNYVHRDLRAANILV	19	0	379	0
	3 of 5	67.60	453	1.62e-05	FPIKWTAPEAA	11	0	427	0
	4 of 5	65.50	822	4.80e-12	SDVWSFGILLTELTTKGRVPYPG	23	0	446	0
	5 of 5	54.36	810	8.07e-12	CPESLHDLMCQCWRKEPEERPTF	23	0	490	0
SH2DOMAIN	1 of 5	56.58	484	2.79e-07	WYFGKITRRESERLL	15	0	151	0
	2 of 5	70.99	392	9.50e-07	RGTFLVRESET	11	0	172	0
	3 of 5	52.83	275	2.39e-06	KGAYCLSVSDFD	12	0	184	0
	4 of 5	59.09	328	1.01e-05	NVKHYKIRKLD	11	0	201	0
	5 of 5	49.75	380	1.37e-07	QFNSLQQLVAYYSKH	15	0	222	0
SH3DOMAIN	1 of 4	49.38	270	2.59e-05	TTFVALYDYES	11	0	87	0
	2 of 4	40.83	352	4.70e-08	TDLSFKKGERLQIVNN	16	0	101	0
	3 of 4	37.35	235	1.60e-02	EGDWWLAHSL	10	0	118	0
	4 of 4	51.24	346	1.43e-06	QTGYIPSNYVAPS	13	0	131	0
SPECTRNALPHA	1 of 3	27.60	322	3.72e-04	VTTFVALYDYESRTETDLSF	20	0	86	0
	2 of 3	40.80	319	3.64e-04	KKGERLQIVNNTEGD	15	0	106	0
GLHYDRLASE15	1 of 7	32.72	199	1.06e-02	GAGGGAFPASQTPSKPAS	18	0	26	0
	7 of 7	26.74	200	3.82e-02	SDVWSFGILLTELTTK	16	0	446	0
ADENOVSFIBRE	3 of 7	26.86	127	9.21e-02	ETDLSFKKGERLQIV	15	0	100	0
	4 of 7	41.30	179	5.47e-03	LRLEVKLGQGC	11	0	270	0
OREXINR	7 of 10	28.57	260	1.65e-02	FGEVWMGTWNGTTRVAIKTLK	21	0	281	0
	10 of 10	32.14	170	7.00e-02	DVWSFGILLTELTT	14	0	447	0
MEROZOITESA	1 of 15	28.95	216	5.32e-02	RAGPLAGGVTTFVALYDYE	19	0	78	0
	11 of 15	20.44	183	7.25e-02	CKVADFGLARLIEDNEYTARQGAKFPIKW	29	0	403	0
PRPRECEPTOR	1 of 10	29.82	180	6.23e-02	KLFGGFNSSDTVTSPQRAG	19	0	62	0
	6 of 10	33.33	180	7.31e-02	AFLQEAQVMKKLRHE	15	0	309	0
GPR153	1 of 9	35.38	184	6.38e-02	SADGHRGPSAAFA	13	0	43	0
	5 of 9	31.58	200	7.63e-02	RLPQLVDMAAQIASGMAYV	19	0	362	0

フィンガープリントを構成するモチーフと対応する配列の表

参考

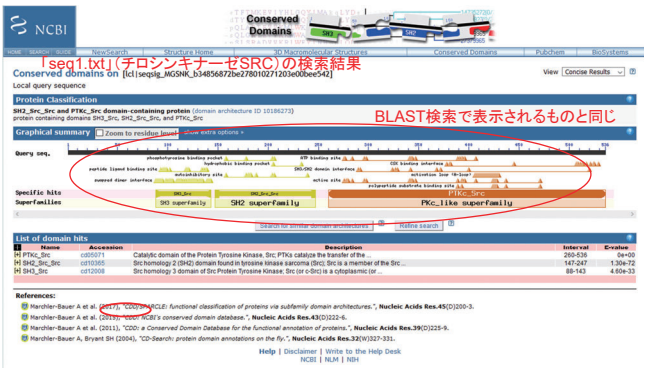
NCBI CDD

- Conserved Domain Database
- タンパク質分類のためのドメインデータベース
- ドメイン情報は、Pfam、SMART、COGなどから取得し、アミノ酸配列に対するBLAST検索の結果と合わせてドメインと配列のリンクを作成
- CD-Search: クエリ配列に対するドメイン検索
- CDART: ドメイン構成（一次元的なつながり）に基づく配列のホモロジー検索
- https://www.ncbi.nlm.nih.gov/cdd



参考

CD-Searchの結果（1）

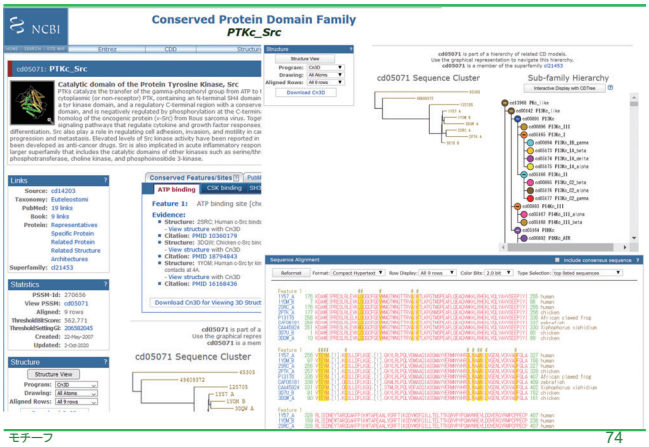


タンパク質の機能情報のレベル

スーパーファミリー	進化的に類縁関係にあるタンパク質のグループ。進化的類縁関係という尺度で最大限に感度を上げてまとめたもの。
クラン（clan）	類似の機能をもつタンパク質のファミリーのグループ。ファミリーをより大きな枠組みで括るのに用いられ、進化は必ずしも関係ない。
ファミリー	進化的な類縁関係が明確で、配列類似度が高いタンパク質のグループ。
サブファミリー	ファミリーを進化的類縁関係、機能により、さらに細かく分類したグループ。ファミリー内の複数の機能を区別して論じるのに用いられる。
ドメイン	タンパク質内の構造的あるいは機能的に独立した部分構造。複数のタンパク質に繰り返し現れる。配列の類似性が見られる。
モチーフ	複数のアミノ酸配列で共通に見られる短い配列の特徴的なパターン。
サイト	高度に保存された短い配列の部分。
シグニチャー（signature）	関連する複数のアミノ酸配列に共通に現れる配列パターンの総称。

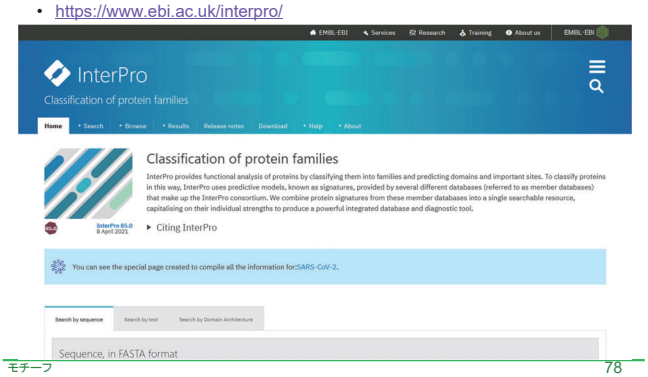
参考

CD-Searchの結果（2）

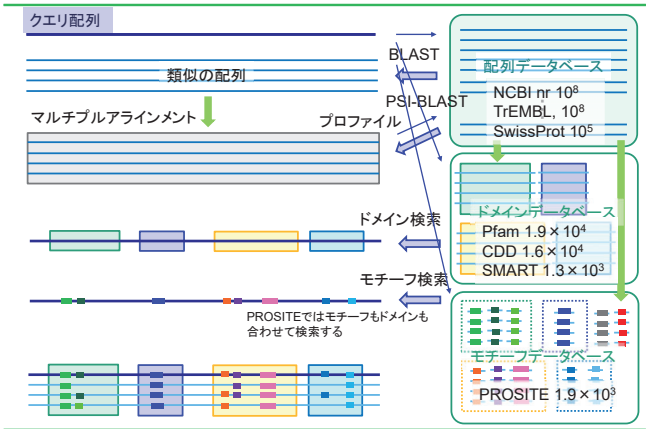


InterPro

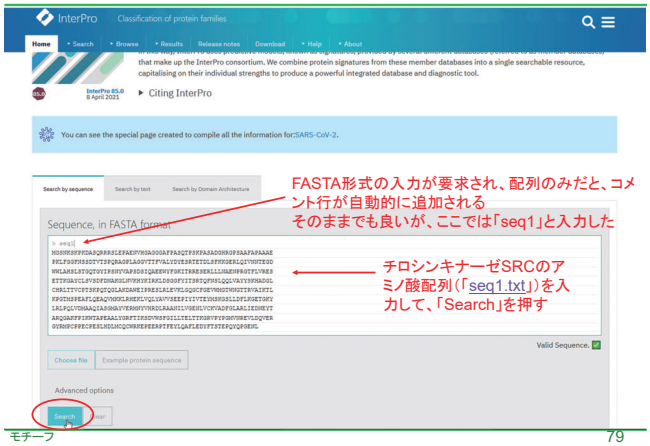
- タンパク質の機能を解析するための総合的なサイト
- タンパク質のファミリー、ドメイン、重要な部位を示す配列特徴（signature）を、複数のデータベースに対してまとめて検索する
- https://www.ebi.ac.uk/interpro/



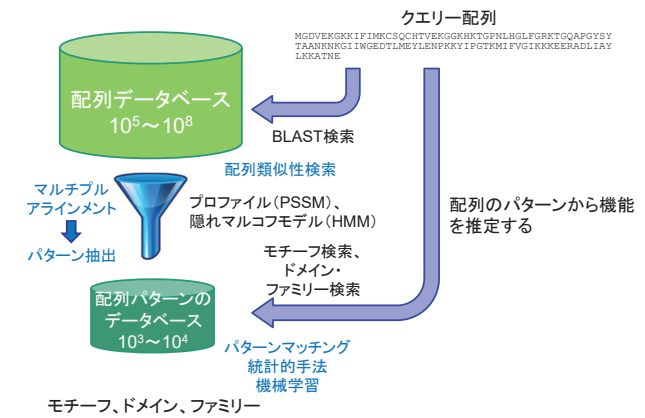
タンパク質の配列からの機能予測



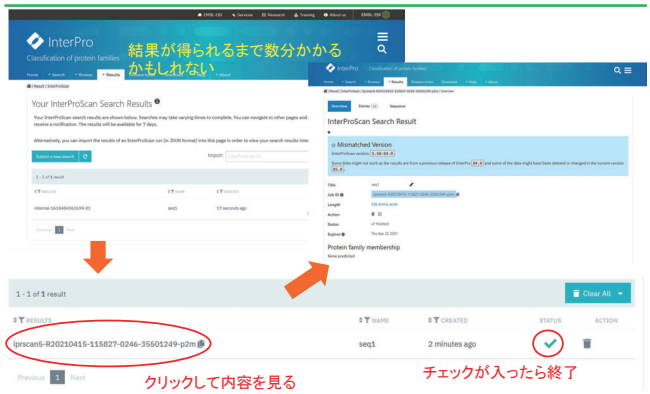
InterProの検索（1）



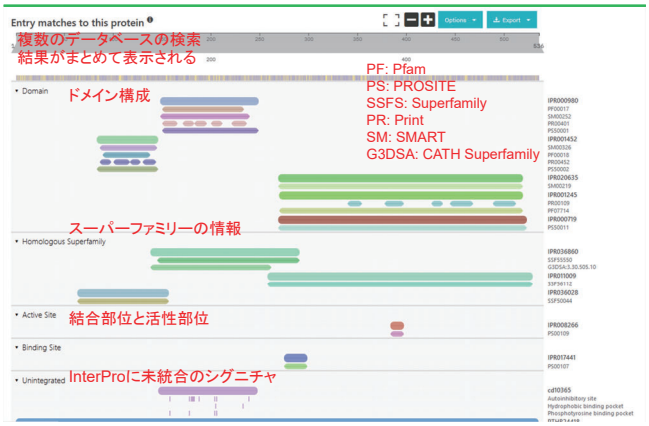
配列から機能を予測する



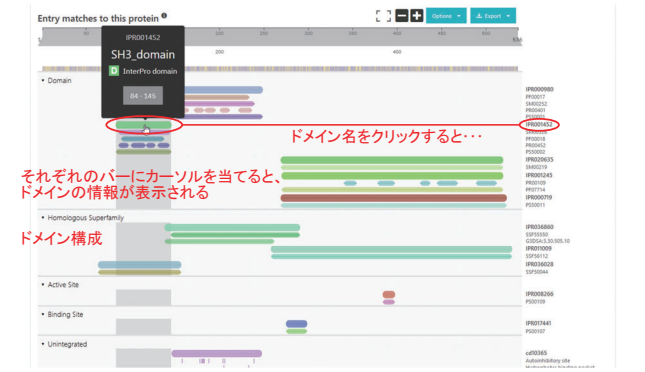
InterProの検索（2）



InterProの検索（3）

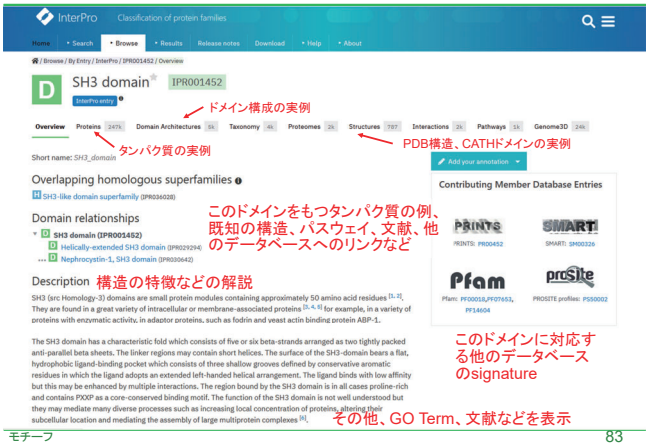


InterProの検索（4）



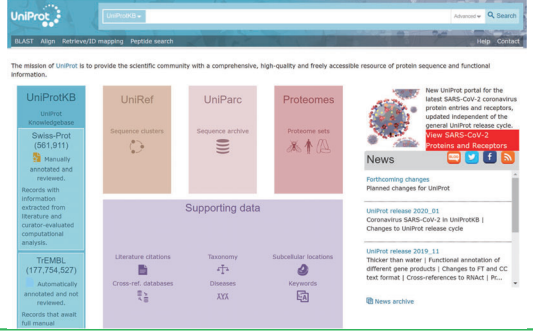
参考

InterProの検索（5）

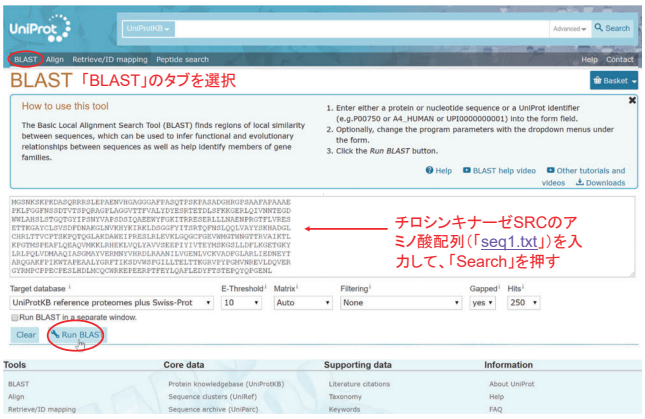


UniProtKB

- UniProt（Universal Protein resource）: EBI、SBI、PIR（Protein Information Resources）が参加するコンソーシアム
- UniProtが提供するアノテーション付きのアミノ酸配列データベース
- https://www.uniprot.org/ KB: Knowledge Base



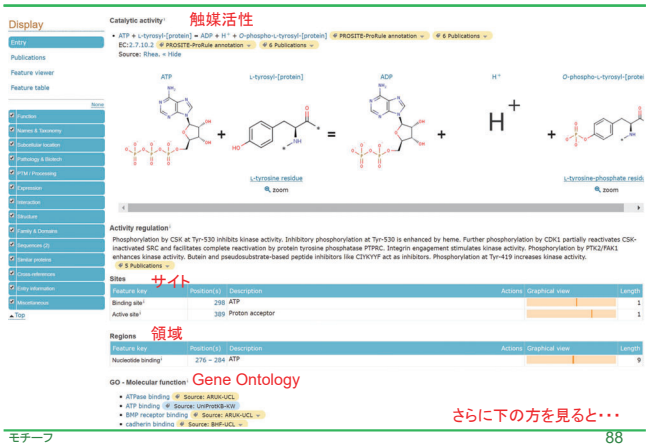
UniProtKBの検索（1）



UniProtKBの検索（2）

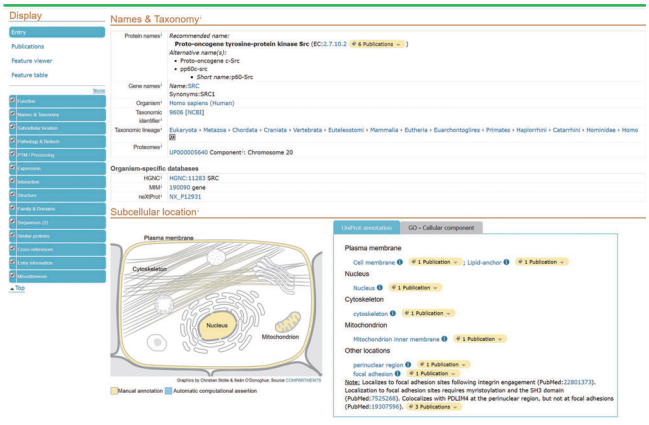
UniProtKBの検索（3）

UniProtKBの検索（4）

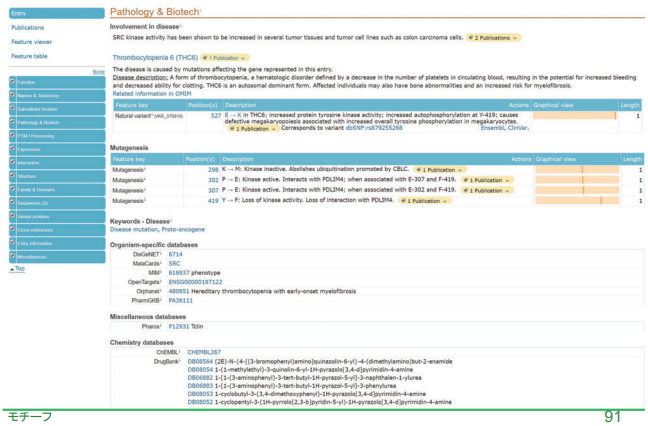


UniProtKBの検索（5）

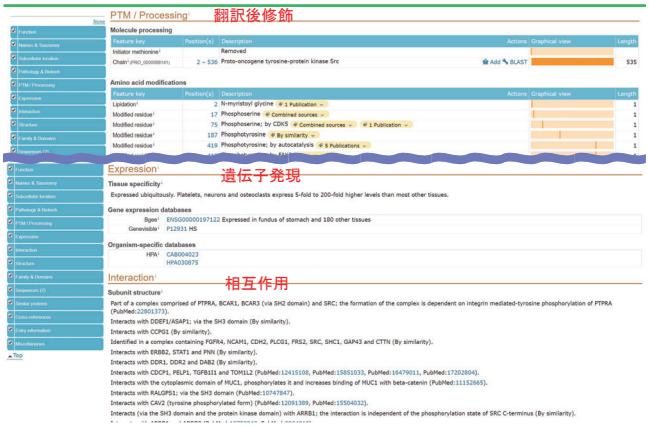
UniProtKBの検索（6）



UniProtKBの検索（7）



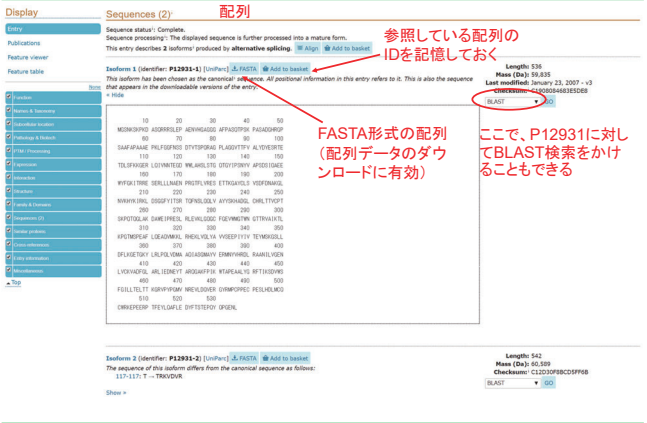
UniProtKBの検索（8）



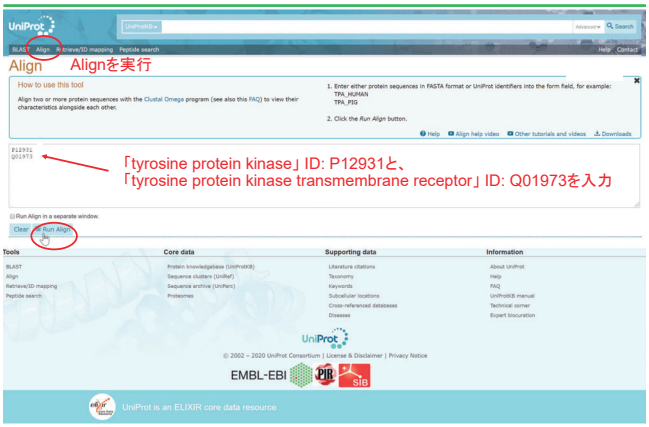
UniProtKBの検索（9）

UniProtKBの検索（10）

UniProtKBの検索（11）



UniProtKBの検索（12）



UniProtKBの検索（13）

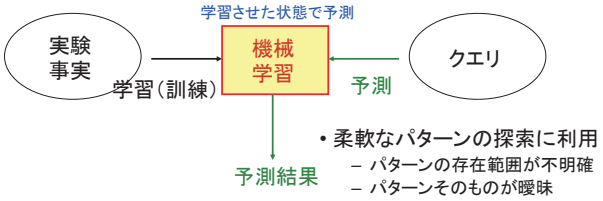
細胞内局在予測とシグナルペプチド予測

- 細胞内局在予測
  - TargetP
    - 真核生物のタンパク質の細胞内局在を予測
    - http://www.cbs.dtu.dk/services/TargetP/
  - PSORT
    - タンパク質の細胞内局在を予測
    - 生物種により異なるツールを用意
    - https://psort.hgc.jp/
- シグナルペプチド予測
  - SignalP
    - シグナルペプチドの存在および切断部位を予測
    - http://www.cbs.dtu.dk/services/SignalP/

ドメイン構成の比較

機械学習の利用

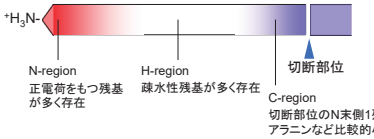
- 機械学習（machine learning）
  - 与えられたデータから知識を習得し、判別や分類を行う
  - 主な手法
    - ニューラルネットワーク
    - HMM（Hidden Markov Model）
    - SVM（Support Vector Machine）
    - ランダムフォレスト
    - 深層学習　など
  - 応用
    - モチーフ検索、二次構造予測、翻訳後修飾予測、相互作用部位予測



シグナル配列

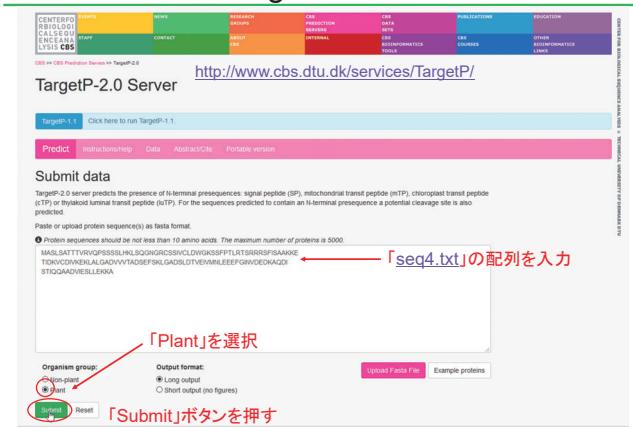
- 細胞内局在、翻訳後修飾（リン酸化、糖鎖付加など）、転写因子結合など、特定の機能に関わる配列部位
- 3～60残基くらいのペプチド配列
- 必ずしも類縁でないタンパク質間で共通に見られる
- 厳密なコンセンサスは必ずしもないが、ある程度の配列特徴がある

例：小胞体輸送シグナル（N末20残基程度）



小胞体内腔に遊離される分泌タンパク質は、シグナルペプチドをもつことで膜を貫通することができる

TargetPの利用

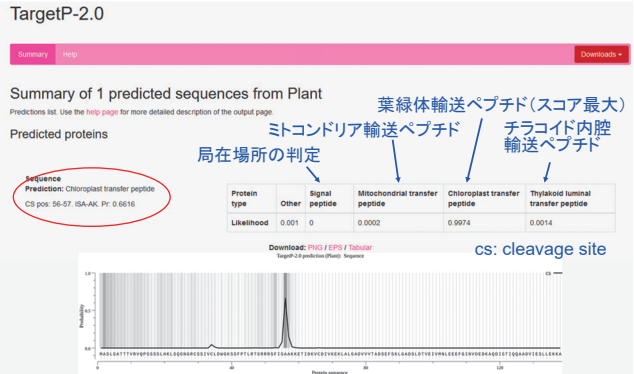


典型的なシグナル配列の例

シグナル配列の機能	シグナル配列の例
核への輸送	-Pro-Pro-Lys-Lys-Lys-Arg-Lys-Val-
核からの輸送	-Leu-Ala-Leu-Lys-Leu-Ala-Gly-Leu-Asp-Ile-
ミトコンドリア内への輸送	$^{+}H_3N$-Met-Leu-Ser-Leu-Arg-Gln-Ser-Ile-Arg-Phe-Phe-Lys-Pro-Ala-Thr-Arg-Thr-Leu-Cys-Ser-Ser-Arg-Tyr-Leu-Leu-
色素体内への輸送	$^{+}H_3N$-Met-Val-Ala-Met-Ala-Met-Ala-Ser-Leu-Gln-Ser-Ser-Met-Ser-Ser-Leu-Ser-Leu-Ser-Ser-Asn-Ser-Phe-Leu-Gly-Gln-Pro-Leu-Ser-Pro-Ile-Thr-Leu-Ser-Pro-Phe-Leu-Gln-Gly-
ペルオキシソーム内への輸送	-Ser-Lys-Leu-COO^-
小胞体内腔への輸送	$^{+}H_3N$-Met-Met-Ser-Phe-Val-Ser-Leu-Leu-Leu-Val-Gly-Ile-Leu-Phe-Trp-Ala-Thr-Glu-Ala-Glu-Gln-Leu-Thr-Lys-Cys-Glu-Val-Phe-Gln-
小胞体内腔に保持	-Lys-Asp-Glu-Leu-COO^-

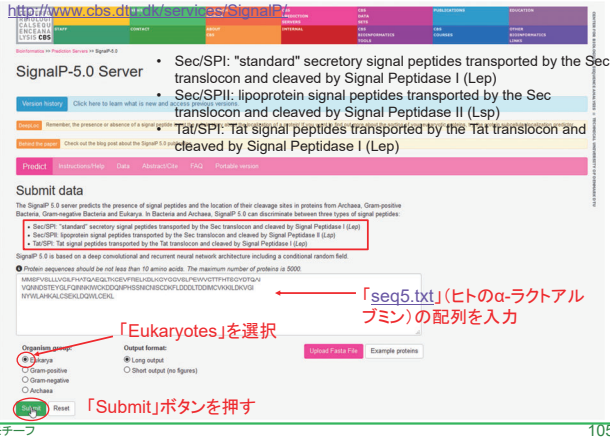
重要な残基は色づけ → 赤: 正電荷、緑: 負電荷、灰色: 疎水性、青: OH基をもつ

TargetPの予測結果の例



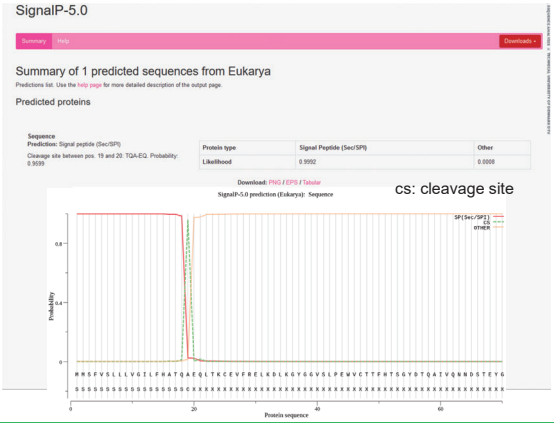
SignalPの利用

http://www.cbs.dtu.dk/services/SignalP/



- Sec/SPI: "standard" secretory signal peptides transported by the Sec translocon and cleaved by Signal Peptidase I (Lep)
- Sec/SPII: lipoprotein signal peptides transported by the Sec translocon and cleaved by Signal Peptidase II (Lsp)
- Tat/SPI: Tat signal peptides transported by the Tat translocon and cleaved by Signal Peptidase I (Lep)

SignalPの予測結果の例



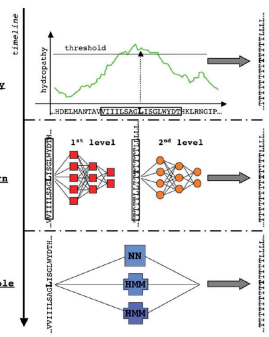
膜タンパク質の性質

- αヘリックス型膜貫通タンパク質の性質と予測
  - 膜貫通領域は疎水性が高いヘリックス領域が多い
    - 20Å程度の疎水性セグメントが細胞膜の30Åの疎水層を貫通
    - ただし、膜タンパク質の25%は、その最も疎水性の高いヘリックスでも、水溶性タンパク質と同程度の疎水性しかもたない
    - 膜貫通領域の候補として、疎水性セグメントを見つける
  - 大きな球状タンパク質は内部に疎水性セグメントをもつ
    - 膜貫通領域と見分ける必要
    - 小さなタンパク質の疎水性セグメントは膜に入りやすい
  - 極性の強い残基(リジン、アルギニン、ヒスチジン、グルタミン酸、グルタミン)、極性が弱くても側鎖の大きな残基(トリプトファン、チロシン)が膜表面付近に存在する傾向がある
    - これらの残基が疎水性の高いヘリックス領域の両端付近に分布しているか調べる

膜タンパク質予測の手法

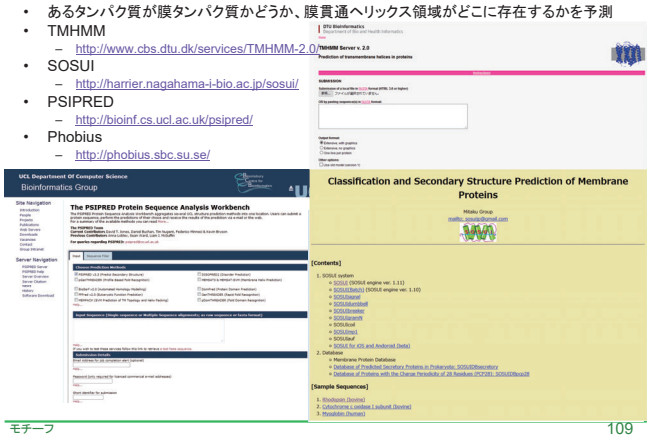
膜貫通領域の予測

- 各残基を中心としたウィンドウごとに疎水性の分布をプロット、一定値を超えた部分を膜貫通領域として予測 **Early**
- 機械学習(ニューラルネットワーク、隠れマルコフモデル、SVMなど)を用いて予測 **Modern**
- 複数の予測器を用いてコンセンサスを予測結果とする **Ensemble**



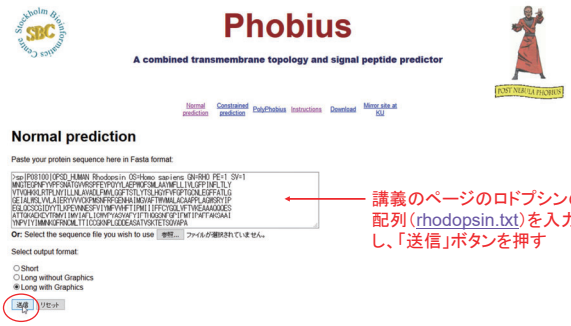
膜タンパク質予測システム

- あるタンパク質が膜タンパク質かどうか、膜貫通ヘリックス領域がどこに存在するかを予測
- TMHMM
  - http://www.cbs.dtu.dk/services/TMHMM-2.0/
- SOSUI
  - http://harrier.nagahama-i-bio.ac.jp/sosui/
- PSIPRED
  - http://bioinf.cs.ucl.ac.uk/psipred/
- Phobius
  - http://phobius.sbc.su.se/

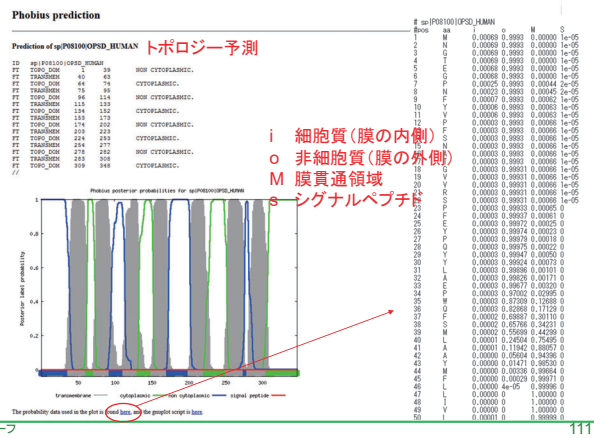


Phobiusの利用

http://phobius.sbc.su.se/



Phobiusの予測結果



課題(1／2)

English version is +3

1. プリンリプレッサータンパク質PurR(UniProtKB ID: P0ACP7(PURR_ECOLI))は、大腸菌のDNA結合タンパク質のlacリプレッサーファミリーのメンバーである。このタンパク質について、以下の問いに答えよ。
   (1) このタンパク質がもつDNA結合に関係するモチーフの名称、そのパターン(正規表現)、その存在位置(残基番号)を答えよ。PROSITEを使用するとよい。
   (2) このタンパク質は、大腸菌のD-ガラクトース結合タンパク質(GBP)と配列が類似していることが知られている。NCBIのBLAST検索のサイトでSwiss-Protデータベースに対してPSI-BLAST検索を行い、そこでヒットしたペリプラズムのGBPのUniProtKB IDを答えよ。また、2つのタンパク質の配列一致度を答えよ。
   (3) (2)において、(1)のモチーフの部分は、GBPにアラインされているか?

   Swiss-Prot は、UniProtKBの一部であり、そのIDはUniProtKB IDで表される。

課題（2／2）

2. 講義のページのファイルseqx.txtのヒトのタンパク質の配列について、以下の問いに答えよ。
   (1) 適当な予測システムを用いて、このタンパク質がシグナルペプチドをもつかもたないかを予測せよ。もつと予測される場合、その位置（残基番号）を求めよ。使用した予測システムを記すこと。
   (2) 適当な予測システムを用いて、このタンパク質が膜貫通領域をもつかもたないかを予測せよ。もつと予測される場合、その位置（残基番号）を求めよ。使用した予測システムを記すこと。
   (3) このタンパク質が登録されている配列データベース（UniProtKBなど）のアノテーション情報と、(1), (2)の予測結果を比較せよ。

出席の確認と課題の提出

- 出席は、ITC-LMSから行って下さい。
  - ワンタイムパスワードは、授業開始時にzoomのチャットでお伝えします。
- 課題の提出は、ITC-LMSから行って下さい。
  - ITC-LMSから提出できない場合は以下のメールアドレスに送って下さい。
    - shimizu@bi.a.u-tokyo.ac.jp
- 授業は20:30に終了します。課題は、そのときまで提出して下さい。
  ~~- 間に合わない場合はzoomのチャットで連絡して下さい。~~
  4月23日まで（24日0時まで）に提出して下さい。
  Submit the assignments until April 23.

Assignments（1 / 2）

1. The purine repressor protein, PurR, is a member of the lac repressor family of Escherichia coli DNA-binding proteins (UniProtKB ID: P0ACP7 (PURR_ECOLI)). Answer the following questions.
   (1) Describe the name, the pattern (regular expression), and the position (in residue number) of the motif which is related to DNA binding contained in PurR. You can use PROSITE.
   (2) This protein is known to have sequences similar to those of E. coli periplasmic D-galactose binding proteins (DBPs). Perform a PSI-BLAST search against Swiss-Prot database in the NCBI BLAST site and describe the UniProtKB ID of the DBP. Also describe the aligned region of the two proteins and the sequence identity of the region.
   (3) Answer whether the above motif (1) is aligned to the DBP of (2) or not when using PSI-BLAST.

Swiss-Prot sequences are identified by the UniProtKB IDs.

Assignments（2 / 2）

2. For the human protein sequence in “seqx.txt”, answer the following questions.
   (1) Is this protein predicted to contain a signal peptide? If so, describe its position (in residue number). Use an appropriate prediction tool.
   (2) Is this protein predicted to contain transmembrane regions? If so, describe their position (in residue number). Use an appropriate prediction tool.
   (3) Compare the annotation described in the sequence database (e.g. UniProtKB) with the prediction results of (1) and (2).

Attendance check and assignments submission

- Attendance is taken via ITC-LMS.
  - You can find one-time password in the zoom chat.
- Submit the assignments to the ITC-LMS site.
  - If you cannot use ITC-LMS, you can submit the assignments to the following address via e-mail.
    - shimizu@bi.a.u-tokyo.ac.jp
- The class ends at 20:30. You should submit the assignments before the end of the class.
  - If you cannot meet the deadline, please contact us via the zoom chat.